Lived Experience Not Found: LLMs Struggle to Align with Experts on Addressing Adverse Drug Reactions from Psychiatric Medication Use

Mohit Chandra[1], **Siddharth Sriraman**[1], **Gaurav Verma**[1], **Harneet Singh Khanuja**[1],
Jose Suarez Campayo[2], **Zihang Li**[3], **Michael L. Birnbaum**[4], **Munmun De Choudhury**[1]
[1]College of Computing, Georgia Institute of Technology
[2]Hospital General Universitario Gregorio Marañón
[3]Hofstra University, [4]Columbia University
`{mchandra9, sidsr, gverma, hkhanuja3}@gatech.edu; jsuarezc@salud.madrid.org`
`zli56@pride.hofstra.edu; michael.birnbaum@nyspi.columbia.edu; munmun.choudhury@cc.gatech.edu`

Abstract

Adverse Drug Reactions (ADRs) from psychiatric medications are the leading cause of hospitalizations among mental health patients. With healthcare systems and online communities facing limitations in resolving ADR-related issues, Large Language Models (LLMs) have the potential to fill this gap. Despite the increasing capabilities of LLMs, past research has not explored their capabilities in detecting ADRs related to psychiatric medications or in providing effective harm reduction strategies. To address this, we introduce the **Psych-ADR** benchmark and the **A**dverse **D**rug Reaction **R**esponse **A**ssessment (**ADRA**) framework to systematically evaluate LLM performance in detecting ADR expressions and delivering expert-aligned mitigation strategies. Our analyses show that LLMs struggle with understanding the nuances of ADRs and differentiating between types of ADRs. While LLMs align with experts in terms of expressed emotions and tone of the text, their responses are more complex, harder to read, and only 70.86% aligned with expert strategies. Furthermore, they provide less actionable advice by a margin of 12.32% on average. Our work provides a comprehensive benchmark and evaluation framework for assessing LLMs in strategy-driven tasks within high-risk domains.

1 Introduction

Adverse Drug Reactions (ADRs)[1] caused by psychiatric medications are a leading cause of hospitalizations among individuals with mental health conditions, accounting for 51.9% to 91.8% of cases, as reported in previous studies (Angadi and Mathur, 2020; Ejeta et al., 2021). With nearly 70% of individuals worldwide having limited access to mental health professionals, many patients increasingly turn to social media platforms such as Reddit to share experiences and seek advice (Kazdin and Rabbitt, 2013; Lee et al., 2017; De Choudhury et al., 2014). Yet, around 35% of posts on mental health-related subreddits go unanswered, leaving many without adequate support (Guimarães et al., 2021). Further, while social media offers a platform for seeking assistance with resolving ADR queries, responses are frequently provided by individuals lacking expertise, raising concerns about the reliability of the information shared (Vosoughi et al., 2018; Wang et al., 2019). Hence, as conversational AI platforms (such as ChatGPT) gain prominence, more individuals are turning to these systems for healthcare-related queries, including those about psychiatric medication and ADRs.

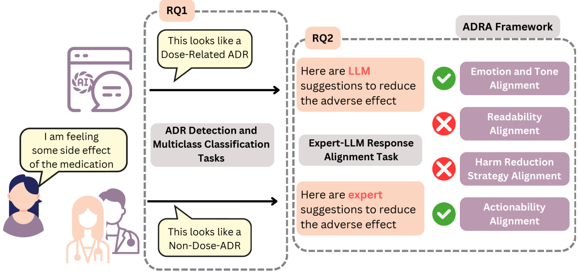

RQ1
This looks like a Dose-Related ADR
I am feeling some side effect of the medication
ADR Detection and Multiclass Classification Tasks
This looks like a Non-Dose-ADR
RQ2
ADRA Framework
Here are LLM suggestions to reduce the adverse effect
Expert-LLM Response Alignment Task
Here are expert suggestions to reduce the adverse effect
✓
✗
✗
✓
Emotion and Tone Alignment
Readability Alignment
Harm Reduction Strategy Alignment
Actionability Alignment

Figure 1: Overview of work; we present two tasks in this work – ADR detection and multiclass classification (**RQ1**), and Expert-LLM response alignment (**RQ2**).

Given the current limitations of healthcare systems and social media platforms, alongside the growing capabilities of LLMs in mental health-related tasks (Yang et al., 2023, 2024; Singhal et al., 2023b), LLMs have the potential to bridge the gap in online discussions by providing high-quality, contextual responses to ADR queries related to psychiatric medications. While previous studies have focused on detecting ADRs using deep learning methods, these efforts have primarily addressed

[1] ADR is defined as *an appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product.* (Edwards and Aronson, 2000).

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11083–11113
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

non-mental health-related ADRs (Mesbah et al., 2019a; Sarker and Gonzalez, 2015; Karimi et al., 2015). Detecting ADRs related to psychiatric medications and evaluation of feedback responses from LLMs towards ADR-related queries, remains unexplored. This gap is particularly significant because addressing ADRs caused by psychiatric medications presents unique challenges, given the complex interplay between mental health conditions, their symptoms, and the potential for psychiatric medications to either alleviate or exacerbate those issues. Furthermore, for LLMs to meaningfully contribute to online discussions and effectively address ADR-related queries, it is essential to rigorously evaluate the quality of their long-form text responses and their alignment with expert knowledge in specialized domains such as psychiatry. This evaluation must consider the LLMs' ability to grasp the complexities of psychiatric medication ADRs and deliver responses that are contextually nuanced. Additionally, there is a need to evaluate the ability of LLMs to portray the lived experiences[2] of healthcare providers in addressing ADR-related queries which has been considered as an important factor in providing effective mental health support. In response to these needs and challenges, we address the following research questions:

RQ1: *How effectively do LLMs detect concerns of ADRs associated with a broad range of psychiatric medications? Additionally, how accurate are LLMs in classifying different types of ADRs?*

RQ2: *To what extent do the responses from LLMs align with those from clinicians across different aspects when addressing ADR queries?*

To answer the above stated research questions, we present **Psych-ADR** benchmark and **A**dverse **D**rug Reaction **R**esponse **A**ssessment (**ADRA**) framework. The proposed Psych-ADR benchmark includes 239 Reddit posts, labeled across two hierarchical levels for ADR detection and multiclass classification along with expert-written responses to queries. The proposed framework evaluates LLM-generated responses against those of medical experts, focusing on four assessment axes: (a) **text readability**, (b) **emotion and tone expression**, (c) **alignment of harm-reduction strategies**, and (d) **actionability of suggested strategies**.

RQ1 results show that both ADR detection and ADR multiclass classification are challenging tasks, with the top model in a few-shot setting achieving F1 scores of 75.38 and 76.69 in respective tasks. We observed that all models exhibited a "risk-averse" nature, leading to a false-positive rate of over 70%. Additionally, models struggled with non-dose-related and time-related ADRs, with GPT-4 Turbo misclassifying 51% and 50% of these instances, highlighting difficulties in grasping nuanced ADR types. For **RQ2**, LLM-generated responses were significantly harder to read than expert-written responses. In contrast, there was no observed significant difference in emotional or tone alignment between LLM and expert responses. However, the best model (OpenBioLLM-70B) achieved only 70.86% alignment with expert harm-reduction strategies, and LLMs provided 12.32% less actionable advice on average. Given the observations, our research has important real-world implications. The proposed benchmark provides a resource for evaluating LLMs on tasks involving the interaction between mental health conditions and psychiatric medications. The proposed framework hold practical utility for policymakers, practitioners, and healthcare professionals to assess LLM performance, especially in strategy-driven tasks in high-risk domains. We have provided code in a repository here[3].

2 Data Collection and Curation

We begin by providing the details of the data collection and filtering pipeline. We used publicly available data in English from Reddit spanning a one-year period (January 2019 - December 2019) obtained from Pushshift Reddit Dataset (Baumgartner et al., 2020). While the broad timeframe ensures large enough data before filtering, the specific period also *(a)* predates the use of generative AI in day-to-day lives and *(b)* the knowledge cutoff for all LLMs used in our evaluation. This allows for a fairer comparison between human experts and LLMs and also ensures minimal presence of machine-generated content on Reddit.

Following the past work (Mesbah et al., 2019b; Saha et al., 2019; Chancellor et al., 2019), we selected 10 subreddits that focus on mental health-related issues or provide a platform for users to ask medical queries (such as r/depression and

[2] *Personal knowledge about the world gained through direct, first-hand involvement in everyday events rather than through representations constructed by other people.*

[3] https://github.com/mohit3011/Lived-Experience-Not-Found

r/askdocs; see Appendix A for the complete list). To extract relevant posts, we compiled a set of 297 FDA-approved psychiatric medications provided by Saha et al. (2019). Further, to detect expressions of adverse symptoms in post titles and texts, we employed HealthE (Gatto et al., 2023), a specialized named entity recognizer for identifying healthcare and medical entities. By combining the psychiatric medication names with the entities given by HealthE, we obtained 19,252 Reddit posts.

Filtering based on mentions of psychiatric medications and adverse symptoms provides a rich sample to extract posts that strictly discuss symptoms caused by psychiatric medications, which is the focus of our study. To specifically filter out posts expressing concerns of adverse drug reactions (ADRs), we prompted GPT-3.5 using definitions and specific conditions identified in previous research while also including insights from co-authors who are medical experts (Edwards and Aronson, 2000); see Appendix B and C for complete list of criteria and exact prompts. Based on the annotations by GPT-3.5, we obtained 6,108 Reddit posts expressing ADR and 11,999 expressing no ADR (rest were deleted). Next, we discuss human validation of the labels for constructing the Psych-ADR benchmark.

3 The Psych-ADR Benchmark

LLM-assisted expert annotations: We conducted expert-led annotations to validate the ADR labels generated by GPT-3.5 and to categorize the specific type of ADR described in each post, if applicable. Given the complexity and time-intensive nature of the human annotation process, we randomly selected 250 posts—consisting of both ADR-labeled and no-ADR-labeled posts as identified by GPT-3.5 for experts to annotate. Based on our discussions with the collaborating medical experts and drawing on the classification provided by Edwards and Aronson (2000), we categorized the ADRs into five granular types– 1) dose-related ADR, 2) non-dose ADR, 3) dose- and time-related ADR, 4) time-related ADR, and 5) withdrawal ADR. We collaborated with three expert annotators — two doctors, and one medical student, all with backgrounds in psychiatry with high proficiency in English. Based on the criteria provided for classifying a post as expressing ADR (Appendix B and C), they annotated each post to determine whether the post described an ADR, and if so, which category of ADR it belonged to along with providing reasoning for it (details related to the annotation tool in Appendix D).

The annotation task proved to be challenging for the annotators, with an average time of ∼7.2 minutes taken to annotate each post due to the complexity and subjectivity inherent in detecting adverse drug reactions. All three annotators agreed on the labels for 48% of the posts. To address the disagreements, we conducted a second round of annotations in which all three annotators collaboratively resolved disagreements, resulting in the final set of labels (details in Appendix D). Finally, 11 posts were discarded due to their lack of relevance to ADRs, resulting in a final benchmark comprising 239 annotated posts. Table 1 presents the statistics for the Psych-ADR benchmark.

Class Label	#Examples
No-Adverse Drug Reaction	106 (44.4%)
Adverse Drug Reaction (ADR)	133 (55.6%)
Non-Dose ADR	93 (38.6%)
Withdrawal ADR	22 (9.2%)
Dose Related ADR	13 (5.4%)
Time Related ADR	4 (1.7%)
Dose and Time Related ADR	1 (0.4%)

Table 1: Class-wise distribution of examples in thePsych-ADR benchmark dataset; % w.r.t. $N = 239$.

Expert responses to ADR posts: A key aspect of Psych-ADR benchmark is the inclusion of expert-written responses to queries in the ADR labeled posts. For each post that expressed an ADR related query, the most experienced annotator (Doctor) provided responses addressing the queries. To facilitate this, we identified and articulated the logical structure of the responses typically seen in clinical settings while working with the medical experts. In accordance to this structure, each response in our dataset begins with empathizing with the patient, followed by information on diagnosis, request for additional information, proposing harm reduction strategies to mitigate the ADR, and concluding with a final set of questions. An example response is shown in Figure 5 in the Appendix.

4 Model Selection & Implementation

We conduct our analysis for the research questions with a total of 9 proprietary and open-weights LLMs. For proprietary models, we evaluate GPT-4o (OpenAI-GPT-4o, 2024), GPT-4 Turbo (Achiam et al., 2023), Claude 3.5 Sonnet, Claude 3 Opus, and Claude 3 Haiku (Anthropic-Claude, 2024). For open-weights models, we

	ADR Detection						ADR Multiclass Classification					
	Zero-Shot		5-Shot-Most Similar		5-Shot-Least Similar		Zero-Shot		5-Shot-Most Similar		5-Shot-Least Similar	
Model	**Acc.**	$\mathbf{F_1}$	**Acc.**	$\mathbf{F_1}$	**Acc.**	$\mathbf{F_1}$	**Acc.**	$\mathbf{F_1}$	**Acc.**	$\mathbf{F_1}$	**Acc.**	$\mathbf{F_1}$
GPT-4 Turbo	72.03	68.55	72.46	69.52	72.46	70.05	**57.58**	**62.16**	65.91	69.36	60.61	64.42
GPT-4o	72.03	69.67	71.19	67.60	69.92	66.71	45.46	47.92	59.85	64.06	58.33	62.23
Llama 3.1-70B Instruct	69.88	65.34	71.97	69.11	71.97	69.29	48.87	52.55	64.66	69.15	62.41	66.88
Llama 3.1-405B Instruct	71.55	68.16	69.88	65.83	73.22	70.91	46.62	50.31	**74.44**	**76.69**	65.41	69.15
Claude 3 Haiku	56.49	42.34	64.02	56.91	65.27	60.74	32.33	34.34	70.68	76.41	54.14	62.00
Claude 3 Opus	**77.41**	**76.44**	**76.57**	**75.38**	**75.73**	**74.39**	42.11	44.68	69.93	73.79	63.16	68.28
Claude 3.5 Sonnet	68.62	63.48	71.13	68.18	**75.73**	73.49	51.13	55.45	70.68	73.87	66.92	**72.66**
OpenBioLLM-Llama3-70B	61.09	50.92	70.71	67.43	69.04	66.17	37.59	37.08	71.43	73.88	67.67	71.02
Llama3-Med42v2-70B	60.25	49.34	64.02	56.53	64.02	56.53	56.40	60.32	72.18	75.16	**68.42**	71.45

Table 2: Performance of different models on Binary Detection and Multiclass Classification tasks under Zero-Shot and 5-Shot scenarios. We report the accuracy score (**Acc.**) and weighted F_1 score as ($\mathbf{F_1}$) with the best and second-best performing model metrics in each scenario highlighted in **bold** and underline, respectively.

evaluate LLama-3.1 405B Instruct-Turbo, LLama-3.1 70B Instruct-Turbo (Meta-LLama3.1, 2024) and specialized medical LLMs – Llama3-Med42-v2 70B (Christophe et al., 2024) and Llama3-OpenBioLLM 70B (Ankit Pal, 2024). The choice of these models stems from their reported performance in different general-purpose and medical benchmarks (Abbas et al., 2024; Nori et al., 2023b; Chen et al., 2023; Anthropic-Claude, 2024).

Previous studies have recommended lower temperatures for detection and labeling tasks to ensure more consistent outputs, while higher temperature values aid in more flexible generation (Jin et al., 2024; Achiam et al., 2023). Accordingly, for the ADR detection and multiclass classification tasks we set the temperature $t = 0$ and use $t = 0.6$ for the response generation tasks. Beyond the task-specific temperature variations, the settings were kept consistent across all the LLMs. Additional details regarding the models, evaluation setup, and compute are provided in Appendix E.

5 RQ1: Detecting Adverse Drug Reaction

For this task, we evaluated LLMs on detecting expressions of adverse drug reactions using the Psych-ADR benchmark. The evaluation involved two separate tasks: (1) identifying the presence or absence of concerns related to ADRs in the 239 Reddit posts, and (2) classifying the type of ADR into one of five pre-defined categories for the 133 instances labeled as expressing ADRs in Psych-ADR benchmark. In both tasks we evaluated models using the zero-shot and few-shot variants of the chain-of-thought (CoT) prompting (Wei et al., 2022). Detailed prompts and classification criteria are provided in Appendix B and C.

Due to the wide variety of medications and symptoms in Psych-ADR benchmark, we evaluated two different example sampling strategies for few-shot prompting. For this, we generated text embeddings for each Reddit title and post using Text-embedding-3-large (OpenAI-TextEmb-3-Large, 2024). Using cosine similarity, we retrieved the five most-similar and five least-similar posts for each example. Table 2 presents the accuracy and weighted F_1 scores for models in the ADR detection and ADR multiclass classification tasks.

5.1 Zero-shot prompting on Psych-ADR

Larger models typically perform better for ADR detection tasks, but this trend does not hold for ADR multiclass classification. As expected, larger models (by parameter size) outperformed their smaller counterparts in the ADR detection task within their respective families, with Claude 3 Opus achieving the highest accuracy at 77.41%, followed by GPT-4o and GPT-4 Turbo at 72.03%. Interestingly, specialized medical models (OpenBioLLM-Llama3-70B and Llama3-Med42v2-70B) struggled in this task. However, for ADR multiclass classification, we did not observe any clear pattern between model size and performance. GPT-4 Turbo was the best performing model with an accuracy of 57.58%, followed by Llama3-Med42v2-70B at 56.40%. All models struggled with multiclass classification, likely due to the complexity of distinguishing between ADR types. Additionally, aligning with prior research, observed results in the multiclass classification showed that larger models do not always excel in specialized tasks (Kanithi et al., 2024).

Models exhibited a "risk-averse" tendency, and prone to commit false-positive errors. In both ADR detection and multiclass classification tasks, all models displayed "risk-averse" behavior, often mislabeling posts without ADRs as positive for

ADRs (see Appendix G for error analysis). In zero-shot settings, Claude 3 Opus had a false-positive rate of 42% for 'ADR-No' labels, while Claude 3 Haiku's false positive rate was as high as 97% (see Appendix F). Similarly, in ADR multiclass classification, models struggled to distinguish between non-dose-related, dose-related, and time-related ADRs. GPT-4 Turbo misclassified 51% of non-dose-related and 50% of time-related ADRs in zero-shot settings. This risk-averse tendency indicates a lack of nuanced understanding of ADR complexities, which could lead to (a) patients discontinuing treatment (Horne and Weinman, 1999; Horne et al., 2005), (b) increased fear about their conditions (Starcevic and Berle, 2013), and (c) "alert-fatigue" among healthcare providers (Phansalkar et al., 2013).

5.2 Few-shot prompting on Psych-ADR

In-context learning enhances model performance but not in every case. We observed the in-context learning in general improved performance of models for both ADR detection and multiclass classification tasks, with a more significant impact on the latter task. For multiclass classification, we observed an average increase of 18.14 and 23.06 points in weighted F_1 score among model performance using least-similar and most-similar example prompting respectively. However, this pattern was not observed in ADR detection task. Claude 3 Opus outperformed other models in the ADR detection, achieving an F_1 score of 76.44 with zero-shot prompting. In ADR multiclass classification, Llama-3.5-405B performed best with most-similar examples (F_1 76.69). For analyzing the impact of providing examples in the ADR detection task, we observed that some models, such as Claude 3 Haiku showed an average improvement of F_1 score (16.49 points), whereas we did not observe such a trend for models such as GPT-4o, Claude 3 Opus in few-shot settings. The stochastic nature of LLM generation, coupled with the inability to learn nuances from examples in the "ADR-No" class, may be a contributing factor to this issue. This was further confirmed as we noted that even in few-shot settings, models exhibited "risk-averse" behavior with high false-positive rates, indicating that providing examples could not effectively compensate for the lack of "lived-experience" in the models. This was the major reason behind models failing to achieve the expected gains in detecting ADR.

Impact of choosing similar or diverse examples depends upon the task. While the performance boost in the ADR multiclass classification task could be attributed to the predominance of non-dose-related ADRs, the comparatively smaller performance gains observed when models were presented with the five least similar examples suggest that models were able to grasp the contextual information presented through the examples and capture the nuances of various ADR types. However, no such pattern was observed in the ADR detection task, with 3 models showing increase in $F_1 \geq 1\%$ with five least similar examples based prompting. This showed that diversity in examples rather than stochasticity impacted model performance.

6 RQ2: Alignment between human and AI feedback

Evaluation of long-form text generation is an open problem and involves many challenges like isolating the stylistics from the semantics. However, in the context of responses to ADR queries, we propose abstracting out the LLM generations and ground-truth expert responses to four key components – (1) emotion and tone, (2) text readability, (3) harm reduction strategy, and (4) actionability of proposed strategies. Via this abstraction to key components, our alignment evaluations focus on specific aspects that contribute towards an ideal response to ADR queries. We explain the importance of these components below and the methodology for evaluation.

Emotional and tone alignment: Emotional intelligence is regarded as a key factor in healthcare, fostering strong therapeutic relationships that drive meaningful change (King Jr, 2011). Therefore, LLM-generated responses should align with expert-written responses in tone and expressed emotion. To assess this, we used Empath (Fast et al., 2016), a widely-used lexicon-based tool, focusing specifically on 8 relevant emotional and tonal categories identified from prior literature (Riess and Kraft-Todd, 2014; Mechanic and Meyer, 2000). We analyzed the distribution of these categories in LLM-generated and expert responses, and quantified their differences using Kullback-Leibler (KL) divergence to measure alignment of expressed emotions and tone in the LLM and expert responses.

Text readability alignment: Past studies have shown that health literacy is strongly correlated with patient outcomes (Wolf et al., 2005). A major factor contributing to lower health literacy is the communication barrier between patients and healthcare providers, which often arises from the complexity of medical text, including the writing style and choice of terminology (DuBay, 2004). Hence, the responses produced by LLMs should be easily readable and be of comparable to that of the expert-written responses. To assess this, we used SMOG index (Mc Laughlin, 1969), a popular readability index to assess health literacy material.

Harm reduction strategy alignment: In cases of adverse reactions to psychiatric medications, suggesting safe medical interventions is crucial to prevent further harm. We operationalized these interventions using *harm reduction strategies* (HRS) (Single, 1995), aimed at minimizing the negative effects of medications that one is reliant on. Ideally, LLMs should propose strategies that align with the expert's responses.

To compare the harm reduction strategies suggested by LLMs and experts, we took inspiration from methods for entailment and factuality evaluation in long-form texts (Min et al., 2023; Wei et al., 2024; Kamoi et al., 2023). First, we extracted atomic HRS from LLM responses by prompting GPT-4o (OpenAI-GPT-4o, 2024). Since some extracted strategies were redundant, we used a few-shot approach to combine those that suggested the same overall approach but differed in specific details to get the final set of HRS for each response (examples in Table 11). To check for the robustness of the extraction and combination method, we conducted a round of human evaluation with 4 annotators. Using a random sample of 40 responses for each task, we evaluated 193 strategies for the extraction and 174 strategies for combination, and obtained a correlation score of 92% and 90% respectively with LLM evaluation.

We then evaluated alignment of HRS for each LLM-expert response pairs using two methods. First, we used AlignScore (Zha et al., 2023), a widely-used unified text alignment method providing a score between 0 and 1 based using a fine-tuned RoBERTa-large model (Liu et al., 2019). We computed AlignScore for each strategy from the LLM response against the expert response. We obtained a response-level AlignScore by averaging the scores across all HRS for the response. Second, for a more interpretable alignment score, we prompted GPT-4o with in-context examples to reason and classify if a strategy is aligned with the expert's response. We computed a response-level GPT-4o score by computing the percentage of aligned HRS over total number of HRS. These two approaches ensured robustness by covering both a continuous alignment score and a binary GPT-4o alignment label. We conducted another round of human evaluation for the GPT-4o score, where four annotators annotated 40 responses, achieving a 95% correlation with GPT-4o's score and reasoning. Prompts for LLM-based tasks are presented in Table 13, 14 & 15, and human evaluation details are presented in Appendix I.1.

Actionability alignment: Prior work in health communication has recognized the importance of *actionability* in the responses of healthcare professionals to enable greater engagement and encourage increased action from patients (Sharma et al., 2023). To this end, we designed an approach to measure the alignment between LLM responses and expert responses along the actionability dimension. We first decomposed actionability into specific sub-dimensions while working with clinical experts and using the guidelines presented in the Patient Education Materials Assessment Tool (PEMAT; AHRQ). Harm reduction strategies recommended by experts and LLMs should be: (i) practical, (ii) contextually relevant, (iii) specific, and (iv) clear. We present concrete definitions for each of the sub-dimensions in Appendix J.

To operationalize the quantification of actionability alignment, we prompted the GPT-4o model using carefully selected in-context learning examples and chain-of-thought prompting. The GPT-4o model considers the ADR post made by the user and assigns a binary label to each harm reductions strategy based on whether or not the target sub-dimension of actionability is present in the strategy (0: absent; 1: present). To validate the labels assigned by the GPT-4o model, the medical experts reviewed the rationales generated for detecting each of the sub-dimensions of actionability in 100 harm reduction strategies, and agreed with 91 of them for practicality, 94 for relevance, 82 and 89 for specificity and clarity, respectively. Overall, the extent of the agreement between experts and GPT-4o rationales reinforced the validity of the labels assigned to the 4 sub-dimensions of actionability. Following this, for responses generated by the LLMs, we computed the fraction of harm reduction strategies that are aligned with the HRS *and* also demonstrate presence of a certain sub-dimension of actionability. For instance, for the practicality dimension, the LLM-generated HRS are scored as:

$$\text{Practicality}_{\text{LLM}} = \frac{\#\text{ aligned \& practical HRS}}{\#\text{ total HRS}}$$

It is worth emphasizing that the constraint of only considering aligned HRS within the LLM-generated responses enforces a penalty for generating unaligned HRS while computing actionability. Since expert responses are inherently always aligned, their HRS do not undergo such a penalization. We present the average scores for the 4 sub-dimensions and their aggregate as the overall actionability score in Table 4.

6.1 Results

Emotional and tone alignment. Figure 2 presents the mean KL-divergence score for the distribution of 8 Empath categories between LLM responses and expert-written response. A χ^2 test was conducted to assess the differences in category distributions, and the p-values were non-significant across all models, indicating that the models' responses were *not* significantly different from the expert-written responses in terms of emotions expressed and the tone used. Further, we observed that larger and more capable models from the Llama and Claude families showed greater alignment with expert responses across different emotional and tone related categories. Interestingly, Llama-3 Med42v2 70B performed the worst. This could be attributed to the fact that a major portion of dataset used for instruction fine-tuning for this model was obtained from the medical and biomedical literature, which may not prioritize emotional communication while providing responses (Christophe et al., 2024).

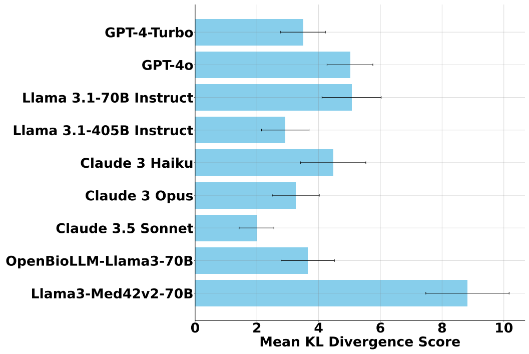


Figure 2: Mean KL Divergence score for the empath categories distribution between models and the expert responses in the Psych-ADR benchmark dataset. (Lower score is better).

Upon closer examination of the individual categories (Figure 8), we found that the expert responses on average showed higher levels of anticipation and affection in the category distribution compared to LLMs. Similarly, a helping tone was more prominent in the expert responses in 6 out of 9 comparisons. However, LLMs exhibited higher use of optimistic and cheerful tones in their responses on average. Additionally, 6 out of 9 LLMs produced responses that used a more polite tone, incorporating more trust-based emotions.

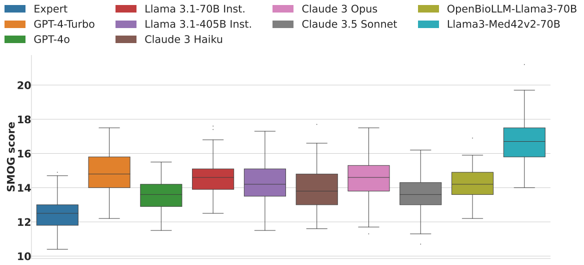


Figure 3: Mean SMOG Scores for Various Models (lower values are better).

Text readability alignment. Figure 3 presents the box plot for LLM and expert response SMOG scores. As observed, LLM-generated responses tend to be more complex, reflected in higher SMOG scores compared to those written by experts (SMOG$_{mean}$12.44) in the Psych-ADR benchmark. Welch's t-test (Welch, 1947) further revealed that the SMOG scores of expert-written responses were significantly lower than those of any LLM-generated responses. We also observed that more capable models produced more readable responses (with Claude 3 Opus being an exception). Similar to the findings on emotional alignment, Llama3-Med42v2-70B showed the lowest alignment with the expert-written responses, producing the most complex responses, likely due to a major portion of instruction-tuning data coming from medical and biomedical scientific literature. In contrast, OpenBioLLM-Llama3-70B outperformed many proprietary models, likely due to the custom dataset used for fine-tuning.

Harm reduction strategy alignment. Table 3 presents the mean response-level AlignScores and GPT-4o scores for alignment of harm reduction strategies of LLMs with the expert's responses. More capable models Llama 3.1-405B Instruct and Claude 3.5 Sonnet in their respective families tended to produce strategies less aligned with the expert than their smaller counterparts, validating a previously observed pattern of LLM performance in responding to open-ended clinical questions (Kanithi et al., 2024). While the open-weights models performed on par or better than proprietary models across both alignment metrics, the best-performing medical model (OpenBioLLM-Llama3-70B) aligned with expert harm reduction strategies for 70.86% of the cases, highlighting the need for further fine-tuning for specialized domains such as psychiatry. Qualitative analysis of non-aligned HRS revealed that most focused on general lifestyle advice, such as maintaining a healthy diet and sleep routine, rather than addressing actions related to the involved medication (details in Appendix I).

	AlignScore		GPT-4o Score	
Model	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**
GPT-4 Turbo	46.49	22.80	65.28	27.22
GPT-4o	42.06	23.10	62.72	27.66
Llama 3.1-70B Instruct	46.91	24.22	63.57	31.50
Llama 3.1-405B Instruct	39.96	20.70	54.71	32.30
Claude 3 Haiku	41.71	25.32	61.96	31.81
Claude 3 Opus	42.42	21.27	59.16	30.21
Claude 3.5 Sonnet	36.83	22.74	59.48	31.63
OpenBioLLM-Llama3-70B	**56.55**	24.81	**70.86**	30.46
Llama3-Med42v2-70B	42.59	22.47	61.61	28.19

Table 3: Alignment of harm reduction strategies of various models with the expert's response. We report the mean and standard deviation for the AlignScore metric GPT-4o score, with the **best** (bold) and second-best (underline) performing model in each metric highlighted.

Actionability alignment. In Table 4, we noted that expert responses scored the highest on overall actionability in comparison to all the LLMs (0.46). Nonetheless, medical models like OpenBioLLM-Llama3-70B and Llama3-Med42v2-70B demonstrate reasonable actionability scores (0.44), followed by other proprietary and open-weights models (0.35 to 0.43). Beyond the aggregate actionability score, the scores for the sub-dimensions provide interesting insights on alignment between expert and LLM responses. While expert responses were rated considerably better than all LLM responses in terms of the practicality (0.83) and contextual relevance (0.73) of the harm reduction strategies, their specificity (0.17) and clarity (0.13) are relatively lacking. This indicates that while LLMs tend to demonstrate greater specificity and clarity in their harm reduction strategy, the recommended strategies may often not be feasible and contextually relevant, considering the users' personal circumstances, such as physical ability, financial resources, and time constraints. This observation further reinforces the need of encoding and reflecting on lived experiences (De Choudhury et al., 2023; Lawrence et al., 2024) as part of ADR responses to address contextual cues, a dimension along which LLMs need to improve further.

Model	Practical	Relevant	Specific	Clear	Actionable
Expert Responses	0.83	0.73	0.17	0.13	0.46
OpenBioLLM-Llama3-70B	0.68	0.70	0.17	0.22	0.44
Llama3-Med42v2-70B	0.60	0.61	0.26	0.29	0.44
Claude 3 Haiku	0.64	0.64	0.21	0.24	0.43
Claude 3.5 Sonnet	0.63	0.61	0.20	0.22	0.42
GPT-4 Turbo	0.63	0.62	0.17	0.21	0.41
Llama 3.1-70B Instruct	0.62	0.64	0.17	0.18	0.40
GPT-4o	0.59	0.57	0.15	0.19	0.38
Claude 3 Opus	0.57	0.54	0.16	0.17	0.36
Llama 3.1-405B Instruct	0.58	0.56	0.13	0.14	0.35

Table 4: Mean actionability alignment scores of HRS (last column), computed as average of practicality, relevance, specificity, and clarity scores.

7 Related Work

Large language models in healthcare: With the growing capabilities of LLMs, past studies have explored their potential to assist stakeholders in healthcare domain. Proprietary models like GPT-4 and MedPalm have shown strong performance on multiple-choice benchmarks and even passed exams such as the USMLE (Singhal et al., 2023a,b; Nori et al., 2023a; Ankit Pal, 2024; Kanithi et al., 2024). LLMs have also been evaluated for mental health support queries (Yang et al., 2023). However, previous research has also highlighted challenges for LLMs in these settings, highlighting cross-lingual disparities (Jin et al., 2024), gender and geographic biases (Restrepo et al., 2024), and limitations in clinical competency tests for both general and mental health (Thirunavukarasu et al., 2023; Jin et al., 2023).

ADR detection and pharmacovigilance: Past research has looked into ADR detection through social media platforms (Mesbah et al., 2019a; Sarker and Gonzalez, 2015; Karimi et al., 2015). However, these studies have predominantly focused on non-mental health related cases, relying on binary classification tasks with limited medication datasets. In contrast, medical studies on ADRs related to psychiatric medications (Angadi and Mathur, 2020; Ejeta et al., 2021) are typically hospital-based, small-scale, and not focused on detecting ADRs within online communities.

Importance of lived experience: Previous research in mental health and psychology has emphasized on the multifaceted importance of lived experience among the patients, educators and healthcare providers. Understanding lived experiences provides insight into individuals' personal realities and preferences, contributing towards a deeper understanding of their experiences, expectations and requirements. In mental health research, previous studies have highlighted the importance of understanding the experiences and realities of individuals living with mental health conditions for providing better treatment (Gilbert and Stickley, 2012; Repper and Carter, 2011). Byrne et al. (2013) further highlighted that students showed positive attitudes and increased self-awareness towards the impact of mental illness on individuals when they were taught by people with lived experience of mental health challenges. Past research in psychology has also stressed on the understanding of the lived experience of individuals belonging to different backgrounds. Previous studies have also highlighted the importance of lived experience in the form of experiential knowledge among the healthcare provider for making decisions (Lyu et al., 2023; Palukka et al., 2021).

8 Conclusion and Future Work

In this work, we proposed the Psych-ADR benchmark and ADRA framework for evaluating the alignment of LLMs with experts on responding ADR queries caused due to psychiatric medication use. In our RQ1 analysis, even the best models achieved only 77.41% accuracy in detecting ADR and 74.44% accuracy in detecting the type of ADR. Our RQ2 analysis further revealed that while models align with experts on expressed emotions and tone of the text, they struggle in important areas like readability, alignment of harm reduction strategies with expert knowledge, and suggesting actionable interventions. Our work can inspire future work to adopt a more holistic approach for evaluating models, emphasizing the integration of “lived experience" alongside expert knowledge.

9 Broader Implications

Responding to ADR queries is challenging due to the complexity of mental health conditions, symptoms, and medication effects. The results from the analyses of RQ1 and RQ2 surface these challenges, revealing nuanced patterns that highlight the intricacies involved. Hence findings from this work have several key implications:

Going beyond the choice-based medical benchmarks. LLMs have achieved near-perfect scores on popular medical benchmarks (Nori et al., 2023a; Singhal et al., 2023b), however, these evaluations typically focus on multiple-choice or case-based questions,which don’t reflect the nuanced understanding required in real-world scenarios like mental health. Despite their strong performance on medical tasks, Llama3-Med42v2-70B and OpenBioLLM-Llama3-70B struggled with detecting ADRs and providing aligned and actionable HRS, highlighting the need to move beyond standard benchmarks towards more holistic alignment evaluation paradigms.

Focusing on empowering experts rather than replacing them. While LLMs did not match expert performance in our analysis, they showed a potential to enhance healthcare by providing clearer, more actionable responses. Given the global shortage of mental health professionals (Kazdin and Rabbitt, 2013), LLMs could expand access to mental healthcare and support experts with further fine-tuning and alignment with expert reasoning.

Disentangling inclusion of humanistic features in LLMs and advocacy for inclusion of lived experience. While our work provides evidence of the lack of lived experience, which is essential for understanding the nuances of a complex task such as ADR detection and for proposing mitigation strategies, we do not advocate for increasing human-like features in LLMs. Previous studies have suggested that heightened anthropomorphism, independent of whether it is accompanied by enhanced capabilities, can increase trust among individuals (Natarajan and Gombolay, 2020; Chen and Park, 2021). Hence, developers and researchers need to be cautious before introducing such features as individuals may trust LLM responses even when they provide incorrect or inconsistent information which can be hazardous in high-risk domains such as healthcare. In contrast, we advocate for approaches that align with previous research, which has shown that the efficacy of LLMs in the healthcare domain can be enhanced through fine-tuning on specialized data or by incorporating useful features into the model, without introducing human-like features (Belyaeva et al., 2023; Li et al., 2023).

10 Limitations

While novel, it is important to acknowledge the limitations of our work. While the proposed Psych-ADR benchmark is the first to focus exclusively on ADRs related to psychiatric medications, the number of examples used for evaluating LLMs are limited, which may not capture the full range of ADRs associated with these medications. Further, we recognize the class-imbalance within the ADR sub-categories, future works can focus on examining better strategies for curating more balanced set of examples. Expanding a benchmark like ours presents challenges due to the time-intensive nature of annotation and response writing, compounded by the subjectivity and complexity inherent in this domain. Additionally, while the responses were provided by a highly experienced doctor, variations in clinical opinions are possible given the subjective nature of ADR assessment in psychiatric medication contexts. Despite these limitations, we believe that the proposed Psych-ADR benchmark provides a valuable resource for further research, offering a robust starting point for the study of ADRs in psychiatric medications.

We also acknowledge certain limitations in the ADRA framework. Although we aimed to compare responses across a set of relevant emotions and tones, our approach relies on a lexicon-based method, which may sometimes miss semantic meaning of the responses. Additionally, the harm reduction strategy alignment in our framework excludes strategies suggested by LLMs that are not present in the expert responses. However, there may be cases where the LLM's proposed strategy is a viable option according to other clinicians, but due to the open-domain nature of the problem and the lack of a verified data source, we were unable to evaluate the correctness of such strategies. Despite these challenges, our work provides a robust framework for assessing the capabilities of LLMs in high-risk strategy-driven domains.

11 Ethical Considerations

We collected public domain social media data from a publicly available dataset which allowed us to use the resource for non-commercial purposes. We further ensured that all data used was de-identified and did not contain any offensive content. As our study involved working with retrospective data without direct interaction with the authors of the posts, the Institutional Review Board (IRB) classified it as non-human subjects research, exempting it from IRB approval. Still, we adhered to established best practices for working with social media data, as recommended in the literature (Weller and Kinder-Kurlanda, 2016, 2015). In line with Reddit's data-sharing guidelines and relevant data-use agreements, we will provide access to the benchmark exclusively comprising Post IDs and annotations, to interested researchers.

Our study presents a systematic approach for evaluating LLMs for addressing ADR related queries from psychiatric medication use, and hence does not inherently pose direct risks. However, it is important to emphasize that better performance on Psych-ADR benchmark should not be interpreted as an indication of increased capabilities in real-world applications. Instead, these results should be complemented with thorough human evaluation to ensure the reliability and safety of the content generated from models.

12 Acknowledgments

Chandra and De Choudhury were partly supported through National Science Foundation (NSF) grant #2230692, a grant from the American Foundation for Suicide Prevention (AFSP). Verma was partly supported by the JPMorgan AI Research PhD and Snap Research Fellowships. Further, this research project has benefitted from the Microsoft Accelerate Foundation Models Research (AFMR) grant program. The findings, interpretations, and conclusions of this paper are those of the authors and do not represent the official views of NSF, AFSP or Microsoft. We further acknowledge the contributions from Santiago Alvarez Lesmes towards data annotation and assessment.

References

Ali Abbas, Mahad S Rehman, and Syed S Rehman. 2024. Comparing the performance of popular large language models on the national board of medical examiners sample questions. *Cureus*, 16(3).

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AHRQ. 2020. The patient education materials assessment tool (pemat) and user's guide. https://www.ahrq.gov/health-literacy/patient-education/pemat9.html. [Accessed 30-08-2024].

Netravathi Basavaraj Angadi and Chhavi Mathur. 2020. Prevalence and severity of adverse drug reactions among patients receiving antipsychotic drugs in a tertiary care hospital. *International Journal of Nutrition,* *Pharmacology, Neurological Diseases*, 10(3):144–148.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Anthropic-Claude. 2024. Introducing Claude 3.5 Sonnet — anthropic.com. https://www.anthropic.com/news/claude-3-5-sonnet. [Accessed 07-09-2024].

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y McLean, and Nicholas A Furlotte. 2023. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pages 86–102. Springer.

Louise Byrne, Brenda Happell, Tony Welch, and Lorna Jane Moxham. 2013. 'things you can't learn from books': Teaching recovery from a lived experience perspective. *International journal of mental health nursing*, 22(3):195–204.

Stevie Chancellor, George Nitzburg, Andrea Hu, Francisco Zampieri, and Munmun De Choudhury. 2019. Discovering alternative treatments for opioid use recovery using social media. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15.

Qian Qian Chen and Hyun Jung Park. 2021. How anthropomorphism affects trust in intelligent personal assistants. *Industrial Management & Data Systems*, 121(12):2722–2737.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms.

Munmun De Choudhury, Meredith Ringel Morris, and Ryen W White. 2014. Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1365–1376.

Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*.

William DuBay. 2004. The principles of readability. *Impact Information*.

I Ralph Edwards and Jeffrey K Aronson. 2000. Adverse drug reactions: definitions, diagnosis, and management. *The lancet*, 356(9237):1255–1259.

Fikadu Ejeta, Temesgen Aferu, Diriba Feyisa, Oliyad Kebede, Jafer Siraj, Workineh Woldeselassie Hammeso, Esayas Tadesse, and Alemayehu Tinishku. 2021. Adverse drug reaction and its predictors among psychiatric patients taking psychotropic medications at the mizan-tepi university teaching hospital. *Neuropsychiatric Disease and Treatment*, pages 3827–3835.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

Joseph Gatto, Parker Seegmiller, Garrett M Johnston, Madhusudan Basak, and Sarah Masud Preum. 2023. Healthe: Recognizing health advice & entities in online health communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1024–1033.

Peter Gilbert and Theodore Stickley. 2012. "wounded healers": the role of lived-experience in mental health education and practice. *The Journal of Mental Health Training, Education and Practice*, 7(1):33–41.

Anna Guimarães, Erisa Terolli, and Gerhard Weikum. 2021. Comparing health forums: User engagement, salient entities, medical detail. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, pages 57–61.

Rob Horne, John Weinman, Nick Barber, Rachel Elliott, Myfanwy Morgan, A Cribb, and I Kellar. 2005. Concordance, adherence and compliance in medicine taking. *London: NCCSDO*, 2005(40):6.

Robert Horne and John Weinman. 1999. Patients' beliefs about prescribed medicines and their role in adherence to treatment in chronic physical illness. *Journal of psychosomatic research*, 47(6):555–567.

Haoan Jin, Siyuan Chen, Mengyue Wu, and Kenny Q Zhu. 2023. Psyeval: A comprehensive large language model evaluation benchmark for mental health. *arXiv preprint arXiv:2311.09189*.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, WWW

'24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Praveen K Kanithi, Clément Christophe, Marco AF Pimentel, Tathagata Raha, Nada Saadi, Hamza Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. 2024. Medic: Towards a comprehensive framework for evaluating llms in clinical applications. *arXiv preprint arXiv:2409.07314*.

Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015. Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR)*, 47(4):1–39.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. *arXiv preprint arXiv:2109.06835*.

Alan E Kazdin and Sarah M Rabbitt. 2013. Novel models for delivering mental health services and reducing the burdens of mental illness. *Clinical Psychological Science*, 1(2):170–191.

Steve H King Jr. 2011. The structure of empathy in social work practice. *Journal of Human Behavior in the Social Environment*, 21(6):679–695.

GG Landis JRKoch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159174.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.

Kathy Lee, Ashequl Qadir, Sadid A Hasan, Vivek Datla, Aaditya Prakash, Joey Liu, and Oladimeji Farri. 2017. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In *Proceedings of the 26th international conference on world wide web*, pages 705–714.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Yang Lyu, Han Yu, Fengli Gao, Xinhua He, and Julia Crilly. 2023. The lived experiences of health care professionals regarding visiting restrictions in the emergency department during the covid-19 pandemic: A multi-perspective qualitative study. *Nursing Open*, 10(5):3243–3252.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

David Mechanic and Sharon Meyer. 2000. Concepts of trust among patients with serious illness. *Social science & medicine*, 51(5):657–668.

Sepideh Mesbah, Jie Yang, Robert-Jan Sips, Manuel Valle Torre, Christoph Lofi, Alessandro Bozzon, and Geert-Jan Houben. 2019a. Training data augmentation for detecting adverse drug reactions in user-generated content. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2349–2359.

Sepideh Mesbah, Jie Yang, Robert-Jan Sips, Manuel Valle Torre, Christoph Lofi, Alessandro Bozzon, and Geert-Jan Houben. 2019b. Training data augmentation for detecting adverse drug reactions in user-generated content. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2349–2359, Hong Kong, China. Association for Computational Linguistics.

Meta-LLama3.1. 2024. Introducing llama 3.1: Our most capable models to date. `https://ai.meta.com/blog/meta-llama-3-1/`. [Accessed 07-09-2024].

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

Manisha Natarajan and Matthew Gombolay. 2020. Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pages 33–42.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023b. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI-GPT-4o. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. [Accessed 07-09-2024].

OpenAI-TextEmb-3-Large. 2024. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/. [Accessed 24-09-2024].

Hannele Palukka, Arja Haapakorpi, Petra Auvinen, and Jaana Parviainen. 2021. Outlining the role of experiential expertise in professional work in health care service co-production. *International Journal of Qualitative Studies on Health and Well-being*, 16(1):1954744.

Shobha Phansalkar, Heleen Van der Sijs, Alisha D Tucker, Amrita A Desai, Douglas S Bell, Jonathan M Teich, Blackford Middleton, and David W Bates. 2013. Drug—drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. *Journal of the American Medical Informatics Association*, 20(3):489–493.

Julie Repper and Tim Carter. 2011. A review of the literature on peer support in mental health services. *Journal of mental health*, 20(4):392–411.

David Restrepo, Chenwei Wu, Constanza Vásquez-Venegas, João Matos, Jack Gallifant, Leo Anthony Celi, Danielle S Bitterman, and Luis Filipe Nakayama. 2024. Analyzing diversity in healthcare llm research: A scientometric perspective. *medRxiv*, pages 2024–06.

Helen Riess and Gordon Kraft-Todd. 2014. Empathy: a tool to enhance nonverbal communication between clinicians and their patients. *Academic Medicine*, 89(8):1108–1112.

Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 440–451.

Punyajoy Saha, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. 2021. "short is the road that leads from fear to hate": Fear speech in indian whatsapp groups. In *Proceedings of the Web conference 2021*, pages 1110–1121.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G Lucas, Adam S Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. *arXiv preprint arXiv:2305.02466*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Eric Single. 1995. Defining harm reduction. *Drug and Alcohol Review*, 14(3):287–290.

Vladan Starcevic and David Berle. 2013. Cyberchondria: towards a better understanding of excessive health-related internet use. *Expert review of neurotherapeutics*, 13(2):205–213.

Arun James Thirunavukarasu, Refaat Hassan, Shathar Mahmood, Rohan Sanghera, Kara Barzangi, Mohanned El Mukashfi, and Sachin Shah. 2023. Trialling a large language model (chatgpt) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Medical Education*, 9(1):e46599.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240:112552.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *Preprint*, arXiv:2403.18802.

Bernard L Welch. 1947. The generalization of 'student's'problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35.

Katrin Weller and Katharina Kinder-Kurlanda. 2015. Uncovering the challenges in collection, sharing and documentation: The hidden data of social media research? In *Proceedings of the International AAAI*

Conference on Web and Social Media, volume 9, pages 28–37.

Katrin Weller and Katharina E Kinder-Kurlanda. 2016. A manifesto for data sharing in social media research. In *Proceedings of the 8th ACM Conference on Web Science*, pages 166–172.

Michael S Wolf, Julie A Gazmararian, and David W Baker. 2005. Health literacy and functional health status among older adults. *Archives of internal medicine*, 165(17):1946–1952.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

A List of subreddits

We use the following list of subreddits to collect data for further filtering: ‘r/depression’, ‘r/anxiety’, ‘r/bipolar’, ‘r/BPD’, ‘r/schizophrenia’, ‘r/autism’, ‘r/mentalhealth’, ‘r/askdocs’, ‘r/diagnoseme’, ‘r/medical_advice’.

The choice for these subreddits stems from past works (Mesbah et al., 2019b; Saha et al., 2019; Chancellor et al., 2019).

B ADR Detection Scenarios and Prompts

Posts on social media platforms discussing adverse drug reactions related to psychiatric medications are often written by individuals with limited or no medical knowledge. As a result, the level of certainty in expressing concerns about potential side effects can vary significantly. Some posts are more assertive, while others express uncertainty. For example, individuals may report experiencing adverse symptoms after taking psychiatric medications, be unsure if these symptoms are caused by the medication, or inquire whether their symptoms could be related to the drugs they are taking. Additionally, some posts may express concerns about possible future side effects of starting a new psychiatric medication. These scenarios were used as examples to guide both annotators and language models. At last, both LLMs and experts were asked to determine whether the concern could be related to ADR or not based on their experience. Table 6 and Table 7 present prompts used with LLMs for detecting cases of ADR from psychiatric medication.

C ADR Multiclass Classification Definitions and Prompts

We provided the same definitions to both the LLMs and expert annotators for the annotation task. To independently evaluate the LLMs, we focused only on posts annotated as expressing ADR-related concerns (N =133) in the Psych-ADR benchmark. Adverse drug reactions (ADRs) related to a psychiatric medications can be classified in one of the following classes:

- **Dose-related**: These are the reactions that are directly related to the dosage of the psychiatric medication.
- **Non-dose-related**: These are the reactions where any exposure of psychiatric medication is enough to trigger an adverse reaction.
- **Time-related**: These are the reactions that are related due to prolonged use in a psychiatric medication which doesn’t tend to accumulate.
- **Dose-and-time-related**: These are the reactions that are related due to dose accumulation, or with prolonged use of the psychiatric medication.
- **Withdrawal**: These are the reactions that are related to the undesired effects of ceasing or stopping the intake of the psychiatric medication.

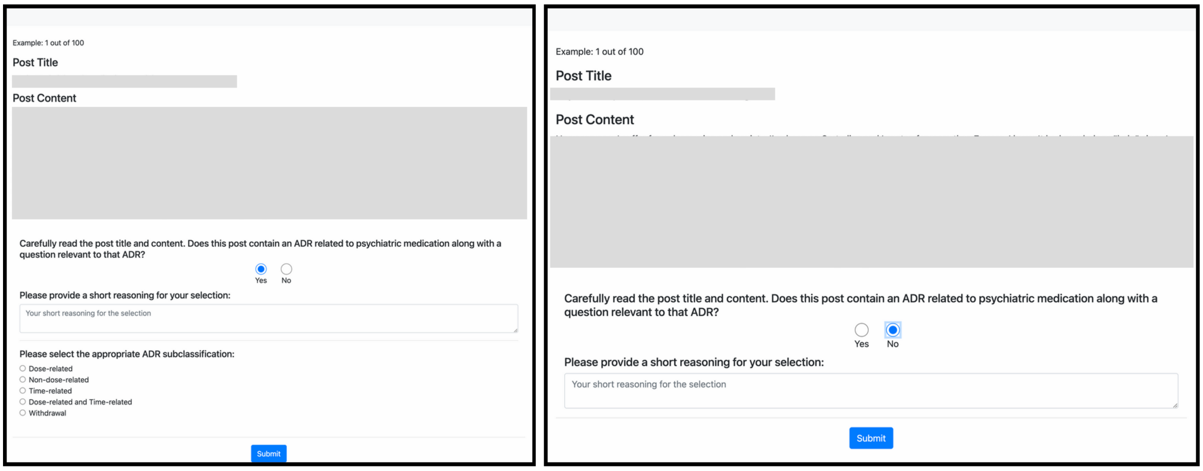
Table 8 and 9 present the LLM prompts used for zero- and few-shot ADR multiclass classification.

D Annotation Task Details

We collaborated with a team of four medical experts (three doctors and one medical student), all of whom are co-authors of this work. Hence, we did not provide any additional compensation for the annotation task. Furthermore, Institutional Review Board (IRB) approval was obtained before the annotation task. To facilitate the annotation process, we developed a custom web-based tool specifically for annotating the Psych-ADR benchmark. Figure 4 presents the interface of the annotation tool used for the data annotation purpose. We conducted a preliminary round of test annotations to familiarize the annotators with both the criteria and the annotation tool. For the further rounds, the average Fleiss’ kappa inter-annotator agreement was $\sim\kappa$ =0.33, with all three annotators agreeing on the labels for 48% of the posts, indicating a fair level of agreement (Landis JRKoch, 1977). These results are consistent with previous research, which have reported similar inter-annotator agreement scores for tasks of comparable difficulty (Karpinska et al., 2021; Saha et al., 2021).

D.1 ADR Post Reply Annotation and Generation

Figure 5 presents the sample of structure of the expert responses provided in the Psych-ADR benchmark. The most experienced doctor on the collaborating team provided the responses, with each taking $\sim$8 minutes to answer on average. We provided the same set of instructions to the domain expert and the LLMs for writing the responses to an ADR-related query. Both the experts and the LLMs were prompted to write responses that follow



Example: 1 out of 100

Post Title

Post Content

Carefully read the post title and content. Does this post contain an ADR related to psychiatric medication along with a question relevant to that ADR?

● Yes ○ No

Please provide a short reasoning for your selection:

Your short reasoning for the selection

Please select the appropriate ADR subclassification:

○ Dose-related
○ Non-dose-related
○ Time-related
○ Dose-related and Time-related
○ Withdrawal

Example: 1 out of 100

Post Title

Post Content

Carefully read the post title and content. Does this post contain an ADR related to psychiatric medication along with a question relevant to that ADR?

○ Yes ● No

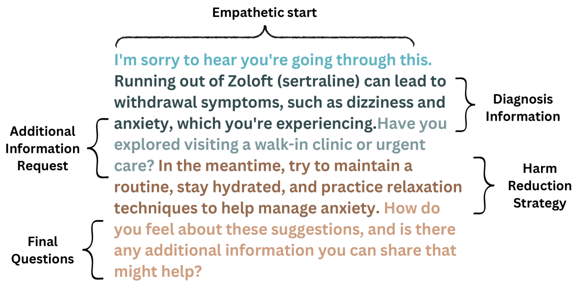
Please provide a short reasoning for your selection:

Your short reasoning for the selection

Figure 4: Annotation interface for the Psych-ADR benchmark used in the annotation process. The interface displays the post title and content, along with access to annotation guidelines. In the left image screenshot, the annotator identifies an adverse drug reaction (ADR) related to psychiatric medication, then provides a brief rationale and selects the class of ADR. In the right image screenshot, the annotator indicates that no ADR is present, in which case only a rationale for this decision is required.

LLM	Version	Parameter Size
GPT-4o	2024-08-06 (OpenAI-GPT-4o, 2024)	(undisclosed)
GPT-4 Turbo	turbo-2024-04-09 (Achiam et al., 2023)	(undisclosed)
Claude 3.5 Sonnet	2024-06-20	(undisclosed)
Claude 3 Opus	2024-02-29	(undisclosed)
Claude 3 Haiku	2024-03-07 (Anthropic-Claude, 2024)	(undisclosed)
LLama-3.1 405B Instruct-Turbo	(Meta-LLama3.1, 2024)	405 billion
LLama-3.1 70B Instruct-Turbo	(Meta-LLama3.1, 2024)	70 billion
Llama3-Med42-v2 70B	(Christophe et al., 2024)	70 billion
Llama3-OpenBioLLM 70B	(Ankit Pal, 2024)	70 billion

Table 5: Models and their corresponding version dates



I'm sorry to hear you're going through this. Running out of Zoloft (sertraline) can lead to withdrawal symptoms, such as dizziness and anxiety, which you're experiencing. Have you explored visiting a walk-in clinic or urgent care? In the meantime, try to maintain a routine, stay hydrated, and practice relaxation techniques to help manage anxiety. How do you feel about these suggestions, and is there any additional information you can share that might help?

Figure 5: Sample answer representing the structure of answers provided in the Psych-ADR benchmark dataset.

a strict logical structure typically seen in clinical settings:

- Start with empathizing with the post writer and acknowledging their concern.
- As step 2, provide a brief explanation for the potential diagnosis as an adverse drug reaction.
- As step 3, talk briefly about the evidence of the adverse drug reaction which is available in the post and provide a brief reasoning for proposing a harm reduction strategy.
- As step 4, ask potential questions to get additional information that might help in providing a better response.
- As step 5, provide a brief explanation of the harm reduction strategy that the poster should follow to get relief.
- Finally, ask the post writer about how they feel about the proposed harm reduction strategy.

Table 10 presents the prompt for ADR reply prompt generation.

Type	Prompt
System Prompt	You are an expert psychiatrist and psychologist, and you help people with their mental health and adverse drug reaction (ADR) related queries. You provide advice, and guidance to people who are experiencing mental health issues. Given below are different events/ways in which concerns related to a potential case of adverse drug reaction (ADR) can be expressed (Presented as ADVERSE_DRUG_REACTION_EVENTS), the Reddit post title (Presented as POST_TITLE), and the Reddit post text (Presented as POST_TEXT) with mentions of psychiatric medicines and symptoms. ADVERSE_DRUG_REACTION_EVENTS: An event can potentially express concerns for an adverse drug reaction (ADR) related to psychiatric medications in one or more of the following four (4) ways: 1) The person is experiencing some adverse symptoms after taking psychiatric medicine(s). 2) The person is unsure if the adverse symptoms are caused due to psychiatric medication(s). 3) The person is inquiring whether the adverse symptoms were potentially caused due to any psychiatric medicine taken by the person. 4) Person is not taking psychiatric medication and is concerned if taking a new psychiatric medication would have side-effects in the future. AND The post asks a question that is relevant to the adverse drug reaction. Keeping the context of the POST_TITLE and POST_TEXT in view and using the different ways of potential expressions provided in ADVERSE_DRUG_REACTION_EVENTS, your task is to determine whether the post actually expresses about an adverse drug reaction (ADR) related to a psychiatric medication and asks a question about the ADR-related to the psychiatric medication or not. You should think step by step and provide a rationale for your answer. You should first provide your rationale and at last you should explicitly provide a label as 'ADR-Yes' or 'ADR-No' determining whether the post talks about adverse drug reaction or not respectively. Exactly provide one of class label and always provide the exact class label in the format - Class Label: <ADR-Yes or ADR-No>
User Prompt	POST_TITLE: <post_title> POST_TEXT: <post_text>

Table 6: Prompt used for the ADR detection task in zero-shot setting.

Type	Prompt
System Prompt	You are an expert psychiatrist and psychologist, and you help people with their mental health and adverse drug reaction (ADR) related queries. You provide advice, and guidance to people who are experiencing mental health issues. Given below are different events/ways in which concerns related to a potential case of adverse drug reaction (ADR) can be expressed (Presented as ADVERSE_DRUG_REACTION_EVENTS), 5 examples of Reddit post title, post text and class label (Presented as EXAMPLE_POST_TITLE, EXAMPLE_POST_TEXT, EXAMPLE_CLASS_LABEL), the Reddit post title (Presented as POST_TITLE), and the Reddit post text (Presented as POST_TEXT) with mentions of psychiatric medicines and symptoms. ADVERSE_DRUG_REACTION_EVENTS: An event can potentially express concerns for an adverse drug reaction (ADR) related to psychiatric medications in one or more of the following four (4) ways: 1) The person is experiencing some adverse symptoms after taking psychiatric medicine(s). 2) The person is unsure if the adverse symptoms are caused due to psychiatric medication(s). 3) The person is inquiring whether the adverse symptoms were potentially caused due to any psychiatric medicine taken by the person. 4) Person is not taking psychiatric medication and is concerned if taking a new psychiatric medication would have side-effects in the future. AND The post asks a question that is relevant to the adverse drug reaction. Set of Examples: 1. EXAMPLE_POST_TITLE: <example_post_title_1> EXAMPLE_POST_TEXT: <example_post_text_1> EXAMPLE_CLASS_LABEL: <example_class_label_1> ... ... EXAMPLE_CLASS_LABEL: <example_class_label_5> Keeping the context of the POST_TITLE and POST_TEXT in view and and using the 5 examples (EXAMPLE_POST_TITLE, EXAMPLE_POST_TEXT, EXAMPLE_CLASS_LABEL) and using the different ways of potential expressions provided in ADVERSE_DRUG_REACTION_EVENTS, your task is to determine whether the post actually expresses about an adverse drug reaction (ADR) related to a psychiatric medication and asks a question about the ADR related to the psychiatric medication or not. You should think step by step and provide a rationale for your answer. You should first provide your rationale and at last you should explicitly provide a label as ‘ADR-Yes’ or ‘ADR-No’ determining whether the post talks about adverse drug reaction or not respectively. Exactly provide one of class label and always provide the exact class label in the format - Class Label: <ADR-Yes or ADR-No>
User Prompt	POST_TITLE: <post_title> POST_TEXT: <post_text>

Table 7: Prompt used for the ADR detection task in 5-shot setting.

Type	Prompt
System Prompt	You are an expert psychiatrist and psychologist, and you help people with their mental health and adverse drug reaction (ADR) related queries. You provide advice, and guidance to people who are experiencing mental health issues. Given below is the list of class names and definitions for each class of adverse drug reaction (ADR) (Presented as ADR_CLASS_NAMES_DEFINITION), the Reddit post title (Presented as POST_TITLE), and the Reddit post text (Presented as POST_TEXT) expressing adverse drug reaction (ADR) related to a psychiatric medication/s. ADR_CLASS_NAMES_DEFINITION: Adverse drug reactions (ADRs) related to a psychiatric medications can be classified in one of the following classes: 1) Dose-related-adr-reactions: These are the reactions that are directly related to the dosage of the psychiatric medication. 2) Non-dose-adr-reactions: These are the reactions where any exposure of psychiatric medication is enough to trigger an adverse reaction. 3) Dose-and-time-adr-reactions: These are the reactions that are related due to dose accumulation, or with prolonged use of the psychiatric medication. 4) Time-related-adr-reactions: These are the reactions that are related due to prolonged use in a psychiatric medication which doesn't tend to accumulate. 5) Withdrawal-adr-reactions: These are the reactions that are related to the undesired effects of ceasing or stopping the intake of the psychiatric medication. Keeping the context of the POST_TITLE and POST_TEXT in view and using the definitions provided in ADR_CLASS_NAMES_DEFINITION, your task is to determine the class of adverse drug reaction (ADR) related to a psychiatric medication/s expressed in the post. You should think step by step and provide a rationale for your answer. You should first provide your rationale and at last you should explicitly provide the class label from ADR_CLASS_NAMES_DEFINITION which is most appropriate and applicable for the post. Only provide one of class label and always provide the exact class label in the format - Class Label: <label name>
User Prompt	POST_TITLE: <post_title> POST_TEXT: <post_text>

Table 8: Prompt used for the ADR multiclass classification task in zero-shot setting.

Type	Prompt
System Prompt	You are an expert psychiatrist and psychologist, and you help people with their mental health and adverse drug reaction (ADR) related queries. You provide advice, and guidance to people who are experiencing mental health issues. Given below is the list of class names and definitions for each class of adverse drug reaction (ADR) (Presented as ADR_CLASS_NAMES_DEFINITION), 2 examples of reddit post title, post text and class label (Presented as EXAMPLE_POST_TITLE, EXAMPLE_POST_TEXT, EXAMPLE_CLASS_LABEL), the Reddit post title (Presented as POST_TITLE), and the Reddit post text (Presented as POST_TEXT) expressing adverse drug reaction (ADR) related to a psychiatric medication/s. ADR_CLASS_NAMES_DEFINITION: Adverse drug reactions (ADRs) related to a psychiatric medications can be classified in one of the following classes: 1) Dose-related-adr-reactions: These are the reactions that are directly related to the dosage of the psychiatric medication. 2) Non-dose-adr-reactions: These are the reactions where any exposure of psychiatric medication is enough to trigger an adverse reaction. 3) Dose-and-time-adr-reactions: These are the reactions that are related due to dose accumulation, or with prolonged use of the psychiatric medication. 4) Time-related-adr-reactions: These are the reactions that are related due to prolonged use in a psychiatric medication which doesn't tend to accumulate. 5) Withdrawal-adr-reactions: These are the reactions that are related to the undesired effects of ceasing or stopping the intake of the psychiatric medication. Set of Examples: 1) EXAMPLE_POST_TITLE: <example_title_1> EXAMPLE_POST_TEXT: <example_text_1> EXAMPLE_CLASS_LABEL: <example_class_label_1> ... EXAMPLE_CLASS_LABEL: <example_class_label_5> Keeping the context of the POST_TITLE and POST_TEXT in view and using the 2 examples (EXAMPLE_POST_TITLE, EXAMPLE_POST_TEXT, EXAMPLE_CLASS_LABEL) and definitions provided in ADR_CLASS_NAMES_DEFINITION, your task is to determine the class of adverse drug reaction (ADR) related to a psychiatric medication/s expressed in the post. You should think step by step and provide a rationale for your answer. You should first provide your rationale and at last you should explicitly provide the class label from ADR_CLASS_NAMES_DEFINITION which is most appropriate and applicable for the post. Only provide one of class label and always provide the exact class label in the format - Class Label: <label name>
User Prompt	POST_TITLE: <post_title> POST_TEXT: <post_text>

Table 9: Prompt used for the ADR multiclass classification task in 5-shot setting.

Type	**Prompt**
System Prompt	You are an expert psychiatrist and psychologist, and you help people with their mental health and adverse drug reaction (ADR) related queries. You provide advice, and guidance to people who are experiencing mental health issues. Given below is the guideline for generating an ideal reply (Presented as IDEAL_REPLY_TEMPLATE), the reddit post title (Presented as POST_TITLE), and the reddit post text (Presented as POST_TEXT) expressing adverse drug reaction (ADR) related to a psychiatric medication/s. IDEAL_REPLY_TEMPLATE: An "ideal reply" follows the below steps: 1. Start with empathizing with the post writer and acknowledging their concern. 2. As step 2, provide a brief explanation for the potential diagnosis as an adverse drug reaction. 3. As step 3, talk briefly about the evidence of the adverse drug reaction which is available in the post and provide a brief reasoning for proposing a harm reduction strategy. 4. As step 4, ask potential questions to get additional information that might help in providing a better response. 5. As step 5, provide a brief explanation of the harm reduction strategy that the poster should follow to get relief. 6. Finally, ask the post writer about how they feel about the proposed harm reduction strategy. Below are some additional guidelines: 1. Do not words like dear poster, dear user, best regards, etc. in the response. 2. Some of the steps of an "ideal answer" are optional and you can skip those if that makes more sense, but it is recommended to include them. 3. The response should be concise and to the point and with the word limit of 225 words or 300 token. Never exceed this word/token limit. Keeping the context of the POST_TITLE and POST_TEXT in view and using the guideline provided in IDEAL_REPLY_TEMPLATE, your task is to generate a reply to the post. Your response should be helpful and aim to provide a solution to the issues/problems mentioned in the post.
User Prompt	POST_TITLE: <post_title> POST_TEXT: <post_text>

Table 10: Prompt used for generating the reply for post expressing ADR related concerns.

E Model Details, Hyperparameters, and Compute

We use API-based model inference for GPT4-Turbo, GPT-4o, Llama 3.1-70B Instruct, Llama 3.1-405B Instruct, Claude 3 Haiku, Claude 3 Opus, Claude 3.5 Sonnet. We used Azure OpenAI service for accessing GPT4-Turbo & GPT-4o models, Together.ai for accessing Llama 3.1-70B Instruct & Llama 3.1-405B Instruct, and the Anthropic platform for accessing the Claude series models. For OpenBioLLM-Llama3-70B and Llama3-Med42v2-70B, we did GPU-based inference using an 8x NVIDIA L40S GPU cluster. The hyperparameters for API-based inference models and GPU-based inference models are presented below. Table 5 presents the details regarding the model sizes and versions.

Hyperparameters for GPT, Llama 3.1 and Claude series models: temperature $t = 0$ (for ADR detection and multiclass classification) & $t = 0.6$ (for response generation), top_p= 1, frequency_penalty= 0, presence_penalty= 0, max_tokens= 600 (for ADR detection and multiclass classification) & max_tokens= 340 (for response generation), number of completions= 1, top_k= 50 (for Claude series models)

Hyperparameters for OpenBioLLM-Llama3-70B and Llama3-Med42v2-70B: temperature $t = 0$ (for ADR detection and multiclass classification) & $t = 0.6$ (for response generation), top_p= 1, frequency_penalty= 0, presence_penalty= 0, max_tokens= 600 (for ADR detection and multiclass classification) & max_tokens= 340 (for response generation), number of completions= 1, top_k= 50.

F ADR detection and multiclass classification results

We analyzed the class-wise distribution of predicted labels for the ADR detection and ADR multiclass classification task. Figure 6 and Figure 7 present the confusion matrices for the ADR detection and ADR multiclass classification task in zero-shot setting. Analyzing Figure 6 we observed that all models performed exceptionally well in cases of ADRs with 4 out of 9 models correctly detecting all examples in the 'ADR-Yes' class. However, all models struggled in correctly classifying cases of 'ADR-No' class with the best model (Claude 3 Opus) misclassifying 42% examples. This qualitatively implied that models showed lack of lived experience and a "risk-averse" behavior. Analyzing the ADR multiclass classification results in zero-shot setting (Figure 7), we observed that *Withdrawal* ADRs were correctly classified more than 90% times by all models. However, all LLMs struggled between the *Dose* and *Non-Dose* related ADRs and failed to understand the nuances between the two types of ADRs.

G ADR Detection and Multiclass Classification Error Analysis

We conducted a qualitative error analysis for misclassified examples in the ADR detection and multiclass classification tasks, focusing on Claude 3 Opus and Llama 3.1 405B in few-shot settings. Upon the analysis, two major themes emerged: (a) lack of lived experience and (b) incorrect assumptions about potential ADR queries. In the first set of errors, models adhered too rigidly to prompt rules, missing other possible symptom explanations. In the second set of errors, models often confused posts seeking emotional support with ADR-related queries. The model misjudged a person's use of social media to share their feelings as a potential ADR-related query.

For cases where the model demonstrates a lack of lived experience, we observe expert quotes such as *"Patient is having swallowing difficulties which seems to be due to GI issues rather than medication"* and *"We can not say she has an ADR since she is actually sleep deprived, plus slightly (minimally) overweight, so we should need to assess if she actually has sleep apnea."*. These quotes indicate that the model is quick to label a post as ADR and can overlook some other contributing factors for the symptoms, while the experts are cautious while labeling a post as ADR. Getting a diagnosis of ADR by psychiatric medication can be overwhelming for patients already suffering from anxiety, depression, and other ailments, and eliminating other potential causes first is a smarter approach.

Reddit is a social media space where people not only ask queries but often share their feelings and thoughts. The model confuses posts of people sharing what they are going through and their experiences as people seeking ADR-related help even if no explicit question has been asked. There are also cases where the question being asked in the post is about the workings of a particular drug,

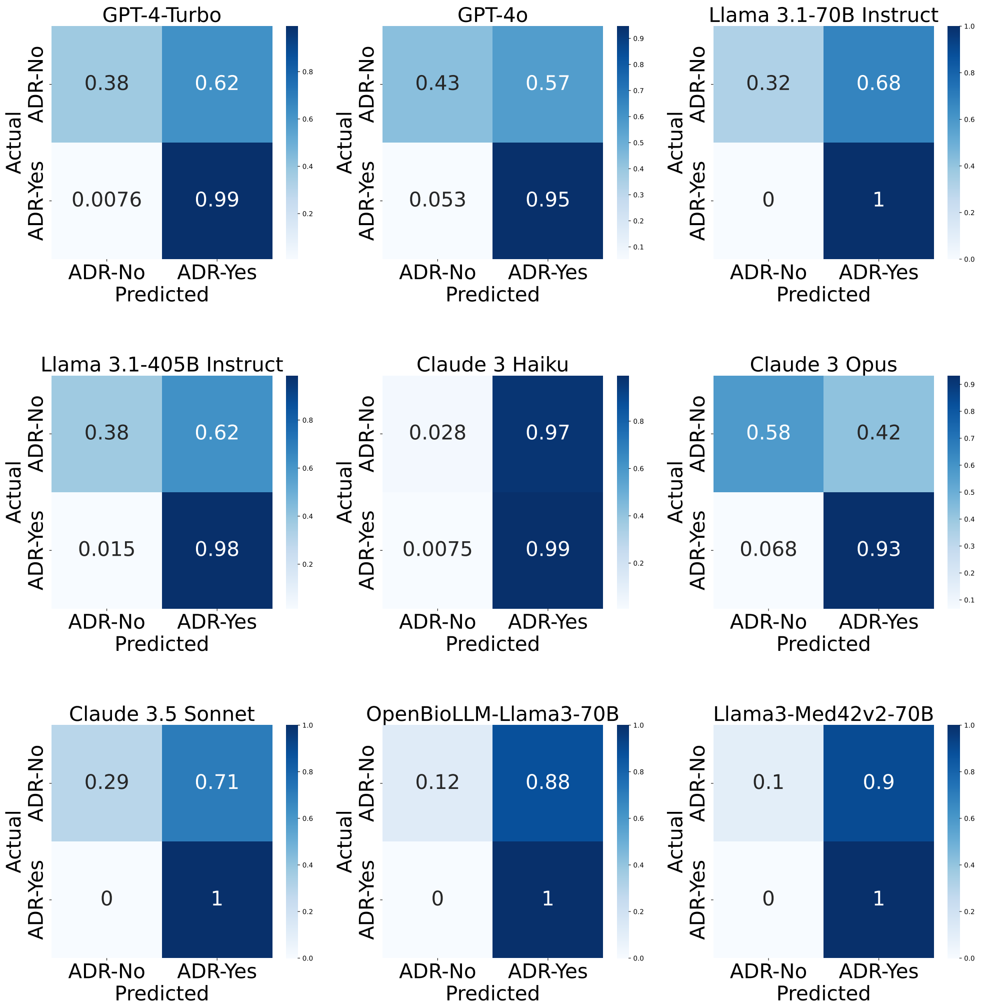


Figure 6: Confusion Matrix for ADR Detection task in zero-shot setting. The values represent the ratio of examples in the predicted class over total number of examples in the actual class.

lifestyle, or something else unrelated to ADR but the model gets confused. Experts are able to identify these correctly and give valid reasoning such as *"Although the post has a detailed description of an ADR (Serotonin syndrome), the patient doesn't have any explicit questions and instead just seems to be sharing her situation."* and *"No ADR, just questions on how the drug works."* for these cases.

H Methodological Details for Emotional and Tone Alignment

To compute the emotional and tonal alignment, we lemmatized the Empath lexicon, expert responses and LLM responses using 'en_core_web_sm' model on SpaCy (Montani et al., 2023). This preprocessing step ensured consistency in comparing the linguistic features across the responses with the Empath categories.

I Harm Reduction Strategy Alignment Qualitative Analysis

We qualitatively analyzed alignment between the LLM's harm reduction strategies and those suggested by the expert. One pattern observed across all LLMs was that in addition to their main response to the issue, they suggested non-pharmacological advice on lifestyle changes involving sleep hygiene (*"Prioritize adequate sleep."*),

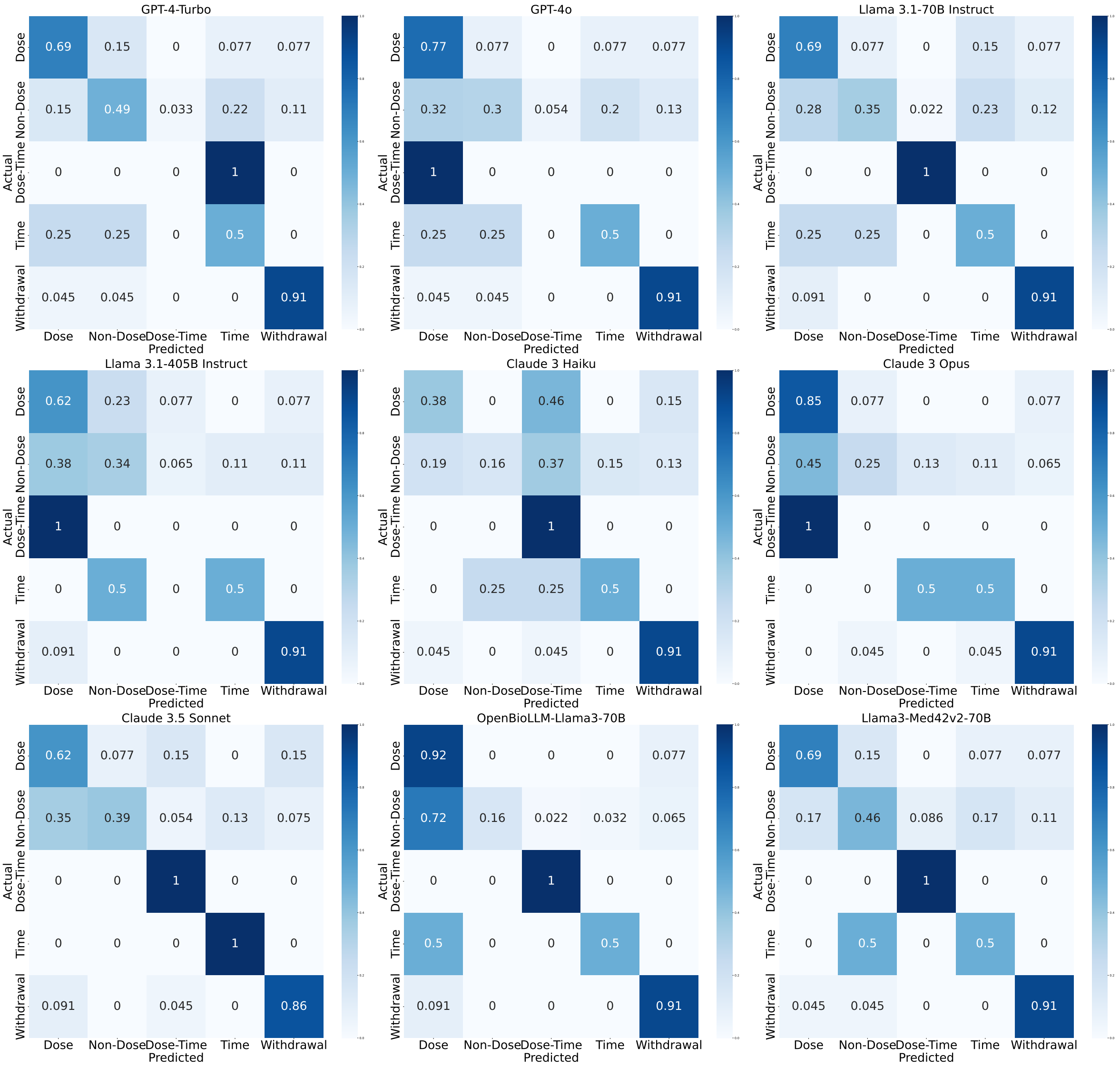


Figure 7: Confusion Matrix for ADR Multiclass classification task in zero-shot setting. The values represents the ratio of number of examples in the predicted class over the total number of examples in the actual (ground truth) class.

diet (*"Maintain a balanced diet."*) and mindfulness techniques such as meditation (*"Practice stress-management techniques like deep breathing."*) and journaling (*"Keep a journal of your symptoms."*). On the other hand, expert answers tended to focus on addressing the symptoms or questions involving the medication. Harm reduction strategies suggested by LLMs related to medication were often paired with an action to discuss with a doctor about the recommendations before committing to them (*"Consider adjusting the dosage with your doctor's guidance."*, *"Gradually taper off Trintellix under medical supervision."*). Examples of alignment in these cases are presented in Table 12.

I.1 Human Evaluation on LLM-based tasks for HRS

For correlation on HRS extraction and alignment between LLMs and humans, we reported the percentage of HRS where the annotator agreed with the LLM's extraction and alignment classification. For combination, since there could be multiple groups formed from different HRS, we reported the percentage of answers where the annotator agreed with the combined HRS that were generated.

J Actionability Criteria

We present the concrete definitions for each of the sub-dimensions of actionability.

- **Practicality:** The proposed strategy should

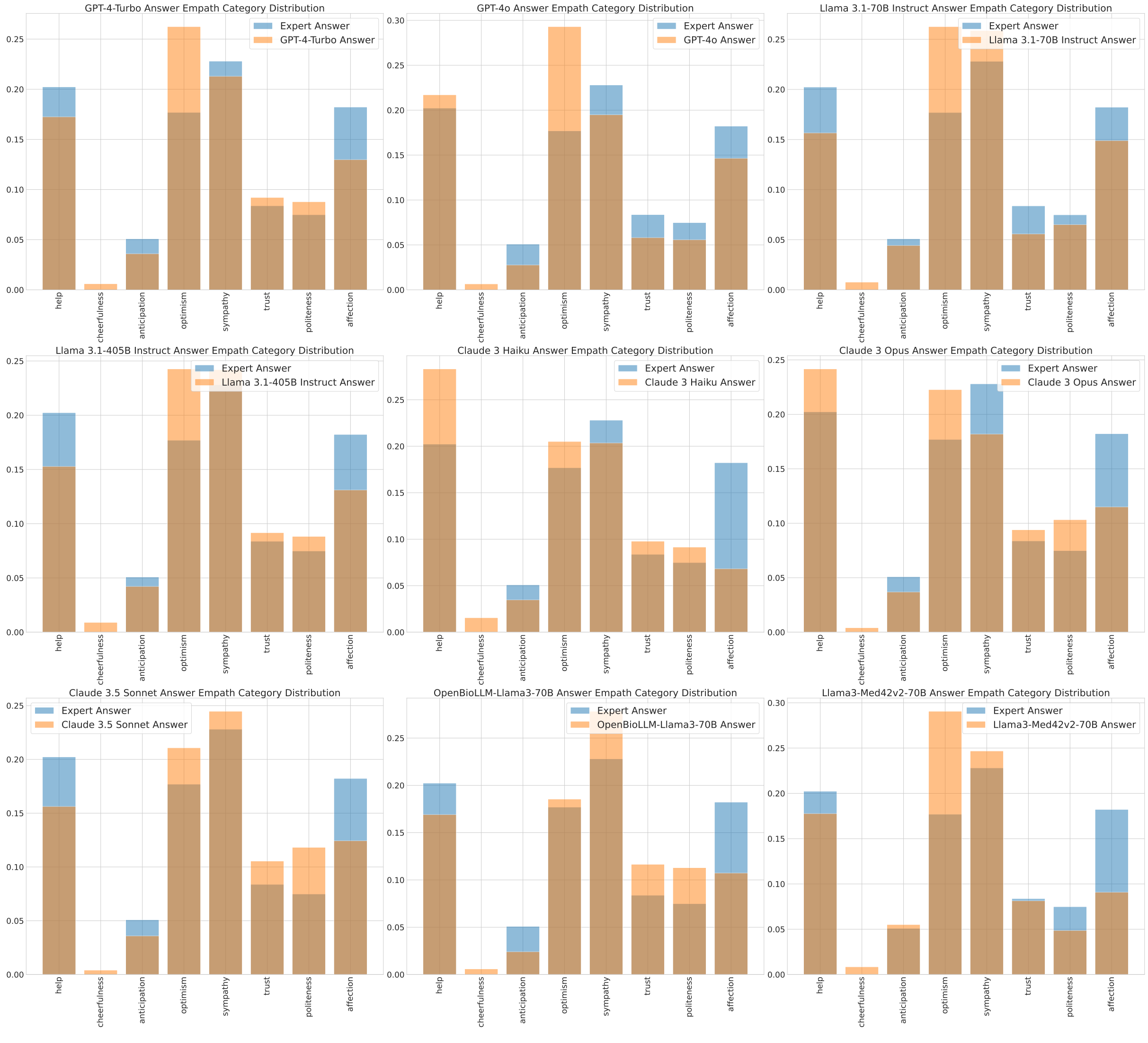


Figure 8: Average Distribution of Emotion and Tone categories from Empath across LLM and Expert responses.

Atomic Strategies	Combined Strategies
1. Consult your healthcare provider before altering your medication dosage. 2. Discuss symptoms and tapering plans with your doctor. 3. Avoid activities that could be dangerous due to vision issues, such as driving.	1. Consult your healthcare provider before altering your medication dosage. Discuss symptoms and tapering plans with your doctor. 2. Avoid activities that could be dangerous due to vision issues, such as driving.
1. Taper off the medication very gradually. 2. Extend the tapering period. 3. Make smaller dosage reductions. 4. Discuss the possibility of using other supportive medications or therapies with your healthcare provider.	1. Taper off the medication very gradually. Extend the tapering period. Make smaller dosage reductions. 2. Discuss the possibility of using other supportive medications or therapies with your healthcare provider.
1. Discuss medication concerns with your psychiatrist. 2. Gradually switch to another antidepressant under supervision. 3. Consider alternative antidepressants like bupropion.	1. Discuss medication concerns with your psychiatrist. 2. Gradually switch to another antidepressant under supervision. Consider alternative antidepressants like bupropion.
1. Review the timing and dosage of medications under the guidance of a psychiatric practitioner. 2. Adjust the time you take Adderall XR. 3. Maintain a sleep routine. 4. Use a sleep mask. 5. Discuss with your psychiatrist the possibility of using a different sleep aid. 6. Consider adjusting the Lamotragine dosage if it's found to be the cause.	1. Review the timing and dosage of medications under the guidance of a psychiatric practitioner. Adjust the time you take Adderall XR. Consider adjusting the Lamotragine dosage if it's found to be the cause. 2. Maintain a sleep routine. Use a sleep mask. 3. Discuss with your psychiatrist the possibility of using a different sleep aid.
1. Switch to another medication if needed. 2. Practice mindfulness techniques. 3. Use relaxation techniques to manage restlessness.	1. Switch to another medication if needed. 2. Practice mindfulness techniques. Use relaxation techniques to manage restlessness.

Table 11: Examples of harm reduction strategies that were combined by GPT-4o (groups highlighted in different colors). The combination was performed for strategies which suggest the same overall approach with minor differences in specific details.

clearly identify at least one action the user can take. Further, it should be contextually feasible/practical, considering their personal circumstances, such as physical ability, financial resources, and time constraints.

- **Contextual relevance:** The provided strategy should be relevant and should contribute to addressing the concern of the patient.

- **Specificity:** The details and instructions provided in the harm reduction strategy should not be vague and should leave no to little room for risky interpretation.

- **Clarity:** The strategy provided can be broken up into clear, manageable and explicit steps.

Table 16 presents the LLM prompts used to get the actionability score for each harm reduction strategy.

K Few-Shot Alignment Evaluation

Table 17 presents the SMOG score, GPT-4o score and overall actionability score for the LLM responses generated using the zero shot and 3-shot setting where the metric values in the zero shot columns correspond to those reported in the alignment tasks presented in the paper. As observed, providing examples of the expert responses did not help in any of the three alignment tasks (readability, harm reduction strategy alignment and actionability alignment). Contrary to the expectation, both the GPT-4o score and overall actionability score decreased in the few shot setting, suggesting that even with the inclusion of answer guidelines and expert-written examples, the model's performance did not improve and, in fact, worsened. On a broader level, these results highlight the need for the use of better training and fine-tuning strategies rather than using post-training methods (such as few-shot prompting) to improve alignment across readability, harm reduction, and actionability. Future works can focus on implementing techniques that capture the

Expert's Response	LLM's Harm Reduction Strategies
"First of all, I understand how difficult living with this experience ... It is always important to make all medication adjustment with close monitorization from your main provider. Therefore, the main recommendation would be to seek for professional help, either by making an appointment with your main provider (if possible) or going to a psychiatry emergency department (if symptoms become severe). There are many strategies for tapering antidepressant medication ..."	1. Consult a healthcare professional familiar with psychiatric medications and withdrawal symptoms. 2. Prioritize adequate sleep. Focus on nutrition. 3. Implement stress management techniques.
"I understand how frustrating living with this feeling might be ... It always depends on each patient's perception and feelings about it. How is your own experience about it? Do you think that it is being helpful or is it worsening your healing process? If you are having difficulties to handle with it we could switch to a different antidepressant to try to avoid this side effect. How do you feel with this? Do you feel more confident with any ... "	1. Discuss with your doctor about adjusting the dosage. 2. Consider alternative medications. 3. Combine medication with therapy, such as cognitive-behavioral therapy.

Table 12: Examples of harm reduction strategies from LLMs and their alignment with the expert's response (aligned strategies are shown in teal, non-aligned strategies in red).

Type	Prompt
System Prompt	A harm reduction strategy is defined as a measure to be taken by an individual to reduce the negative effects of consuming a psychiatric medication. This could include changing the dosage (frequency or time of taking it) of a medication, doing exercises, avoiding certain food items, taking alternative medication or treatment, consulting a healthcare provider etc. Instructions: 1. You are given a RESPONSE from a health expert. Your task is to extract as a list of atomic harm reduction strategies from the RESPONSE. 2. An atomic harm reduction strategy should contain an action verb and contain a single piece of advice. 3. An atomic harm reduction strategy should be extracted from a statement in the RESPONSE and not from a question. 4. Each atomic harm reduction strategy should carry an entirely different piece of advice, and should be independent of other atomic harm reduction strategies in the list. 5. You should only output the atomic harm reduction strategies as a list, with each item starting with "- ". Do not include other formatting.
User Prompt	RESPONSE: <response>

Table 13: Prompt used for the harm reduction strategy extraction task.

nuances of mental health related conditions and use a context-aware approach for better alignment with medical experts. Finally, given the worsened performance in the few-shot setting, we report only the zero-shot results in the paper, which in turn highlights the misalignment between LLMs and expert responses.

Type	Prompt
System Prompt	A harm reduction strategy is defined as a measure to be taken by an individual to reduce the negative effects of consuming a psychiatric medication. This could include changing the dosage (frequency or time of taking it) of a medication, doing exercises, avoiding certain food items, taking alternative medication or treatment, consulting a healthcare provider etc. Instructions: 1. You are given a list of harm reduction strategies from a health expert. 2. Your task is to combine similar harm reduction strategies by logically grouping them based on the similarity of the advice. 3. Two harm reduction strategies are similar if they suggest the same overall approach with differences only in the specific details. 4. Do not alter the wording of any of the harm reduction strategies, only group them as multiple sentences in a single combined harm reduction strategy. 5. You should only output the combined harm reduction strategies as a list, with each item starting with "- ". Do not include other formatting. You should combine strategies based on groups such as: 1. Lifestyle changes such as sleeping patterns, diet, physical exercise. 2. Mindfulness-based exercises. 3. Adjusting dosage of existing medication. 4. Trying out new medications. 5. Consulting people for different opinions. You should NOT: 1. Combine strategies purely based on the person involved in the suggestion (e.g: doctor). 2. Combine strategies that suggest full-fledged therapy approaches with those that suggest simple self-imposed mindfulness exercises. Do this for the harm reduction strategies under "Your Task:". Consider the following examples: Harm Reduction Strategies: <example_1_harm_reduction_strategies_list> Reasoning: <example_1_reasoning> Combined Harm Reduction Strategies: <example_1_combined_harm_reduction_strategies_list> ... ... <example_5_combined_harm_reduction_strategies_list>
User Prompt	Your Task: Harm Reduction Strategies: <harm_reduction_strategies_list>

Table 14: Prompt used for the harm reduction strategy combination task in 5-shot setting.

Type	Prompt
System Prompt	You are an intelligent agent who is given a RESPONSE from a psychiatrist to a patient, and a LIST OF STATEMENTS that are harm reduction strategies. A STATEMENT is considered 'Suggestion-Present' if it can be broadly inferred implicitly OR explicitly as a harm reduction strategy suggested in the RESPONSE, or else it is 'Suggestion-NotPresent'. A 'Suggestion-Present' STATEMENT can be a specific instantiation of a broad harm reduction strategy mentioned in the RESPONSE or vice-versa. The RESPONSE may contain generic names of medication, while a STATEMENT may use a brand name for the same medication, note that these are considered the SAME. Instructions: 1. The following LIST OF STATEMENTS is related to the context of the given RESPONSE. 2. Your task is to analyze if EACH STATEMENT is considered 'Suggestion-Present' or 'Suggestion-NotPresent', based on the given definition and the RESPONSE. 3. One by one, for each STATEMENT, mention step-by-step reasoning behind the classification, along with the label. The reasoning and classification for each STATEMENT should be independent of other STATEMENTS. 4. After doing this for each STATEMENT, state the total NUMBER of 'Suggestion-Present' STATEMENTS, in a new line starting with "Number of 'Suggestion-Present' statements in total:". 5. Answer ONLY in plain text (without Markdown formatting) for the RESPONSE and LIST OF STATEMENTS under "Your Task". Consider the following examples: RESPONSE: <example_1_response> LIST OF STATEMENTS: <example_1_harm_reduction_strategies_list> SOLUTION: <example_1_solution> ... ... <example_5_solution>
User Prompt	Your Task: RESPONSE: <response> LIST OF STATEMENTS: <harm_reduction_strategies_list>

Table 15: Prompt used for the harm reduction strategy alignment task in 5-shot setting.

Type	Prompt
System Prompt	You are an expert psychiatrist and psychologist. You are also an expert in identifying the practicality, contextual relevance, specificity, and clarity of harm reduction strategies. You will be provided with the original query posed by the health advice seeker (presented as ORIGINAL_QUERY) and asked to the detect whether or not the harm reduction strategy suggested by the healthcare provider (presented as HARM_REDUCTION_STRATEGY) meets the criteria for being practical, contextually relevant, specific, and clear. We define each of the dimensions as follows: Practicality: The proposed strategy should clearly identify at least one action the user can take. Further, it should be contextually feasible/practical, considering their personal circumstances, such as physical ability, financial resources, and time constraints. Contextual relevance: The provided strategy should be relevant and should contribute to addressing the concern of the patient. Specificity: The details and instructions provided in the harm reduction strategy should not be vague and should leave no to little room for risky interpretation. Clarity: The strategy provided can be broken up into clear, manageable and explicit steps. Make sure that you output the results strictly in the following format: {’rationale_to_assess_practicality’: ’<your rationale goes here>’, ’practicality_decision’: ’<0 or 1>’, ’rationale_to_assess_contextual_relevance’: ’<your rationale goes here>’, ’contextual_relevance_decision’: ’<0 or 1>’, ’rationale_to_assess_specificity’: ’<your rationale goes here>’, ’specificity_decision’: ’<0 or 1>’, ’rationale_to_assess_clarity’: ’<your rationale goes here>’, ’clarity_decision’: ’<0 or 1>’} where 0 indicates that the strategy does not meet the criteria and 1 indicates that it does. Consider the following examples: ORIGINAL_QUERY: <example_1_query> HARM_REDUCTION_STRATEGY: <example_1_hrs> OUTPUT: {’rationale_to_assess_practicality’: <example_1_practicality_rationale>, ’practicality_decision’: <example_1_practicality_decision>, ’rationale_to_contextual_relevance’:<example_1_contextual_relevance>, ’contextual_relevance_decision’: <example_1_contextual_relevance_decision>, ’rationale_to_assess_specificity’:<example_1_specificity_rationale>,’specificity_decision’: <example_1_specificity_decision>, ’rationale_to_assess_clarity’:<example_1_clarity_rationale>,’clarity_decision’:<example_1_clarity_decision>} ... ... ’rationale_to_assess_clarity’:<example_3_clarity_rationale>,’clarity_decision’:<example_3_clarity_decision>}
User Prompt	ORIGINAL_QUERY: <query> HARM_REDUCTION_STRATEGY: <harm_reduction_strategy>

Table 16: Prompt used for the decomposition of actionability in harm reduction strategies.

Model Name	**SMOG Score (Zero Shot)**	**SMOG Score (3-Shot)**	**GPT-4o Score (Zero Shot)**	**GPT-4o Score (3-Shot)**	**Overall Actionability Score (Zero Shot)**	**Overall Actionability Score (3-Shot)**
GPT-4 Turbo	14.83	14.70	65.28	59.68	0.41	0.40
GPT-4o	13.50	14.13	62.72	57.68	0.38	0.34
Llama 3.1-70B Instruct	14.57	13.97	63.57	59.17	0.40	0.38

Table 17: Comparison of model performance across SMOG score, GPT-4o score and overall actionability score for the LLM responses generated using the zero shot and 3-shot setting where the metric values in the zero shot columns correspond to those reported in the alignment tasks presented in the main paper.